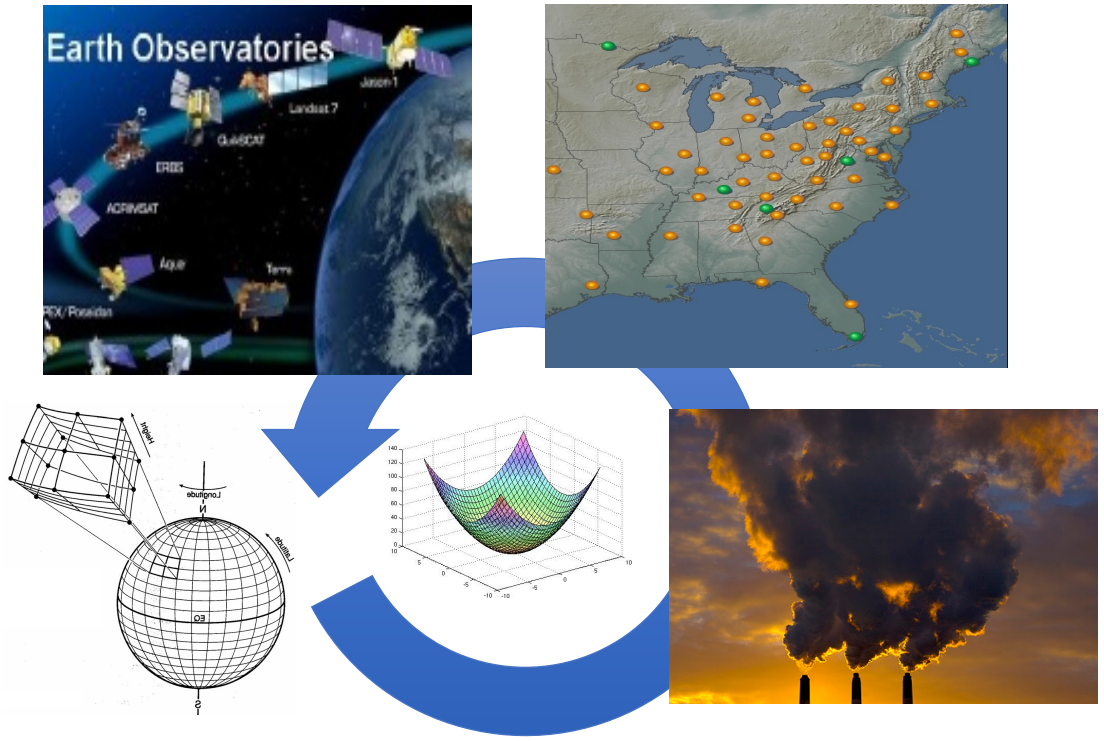
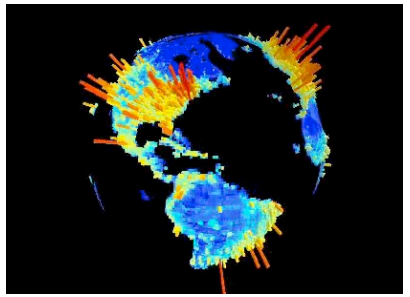
Surrogate Modeling for Atmospheric Chemistry and Data Assimilation



Daven K. Henze, Alireza Doostan (PIs), William Tsui, Hee-sun Choi (Postdocs), CU Boulder
Nicolas Bousserez, ECMWF
Support from NASA AIST

Air quality modeling

Ambient air quality leads to millions of premature deaths annual (Cohen et la., 2017).



The same species leading to air pollution also impact climate change (some cooling, some warming), or are co-emitted with large sources of long-lived greenhouse gases (e.g., CO_2, SO_2, and NO_x from power plants).

Air quality and climate models are used to evaluate the impacts of environmental policy and understand atmospheric processes.

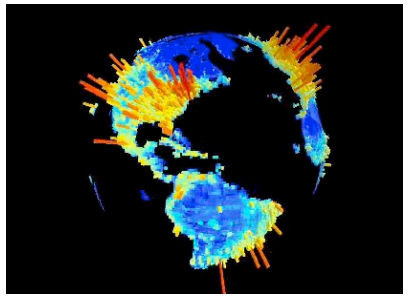
Air quality modeling

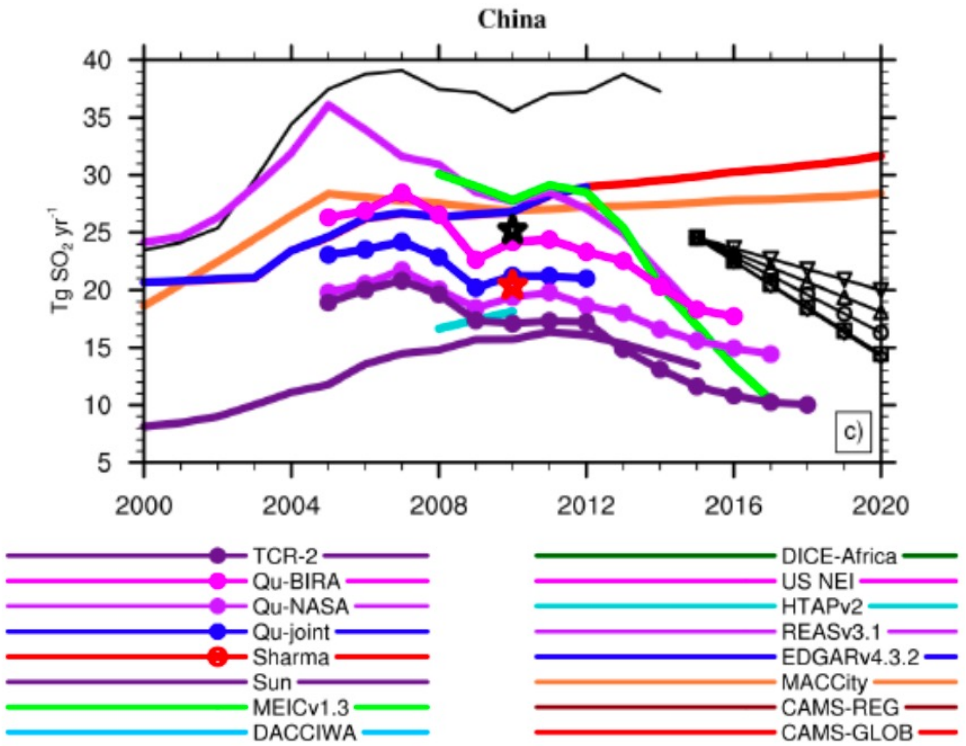
Ambient air quality leads to millions of premature deaths annual (Cohen et la., 2017).

The same species leading to air pollution also impact climate change (some cooling, some warming), or are co-emitted with large sources of long-lived greenhouse gases (e.g., CO_2, SO_2, and NO_x from power plants).

Air quality and climate models are used to evaluate the impacts of environmental policy and understand atmospheric processes.

Our understanding of the magnitude of the inputs to these models, the emissions, are often accurate to 20% at best, 1000% at worst

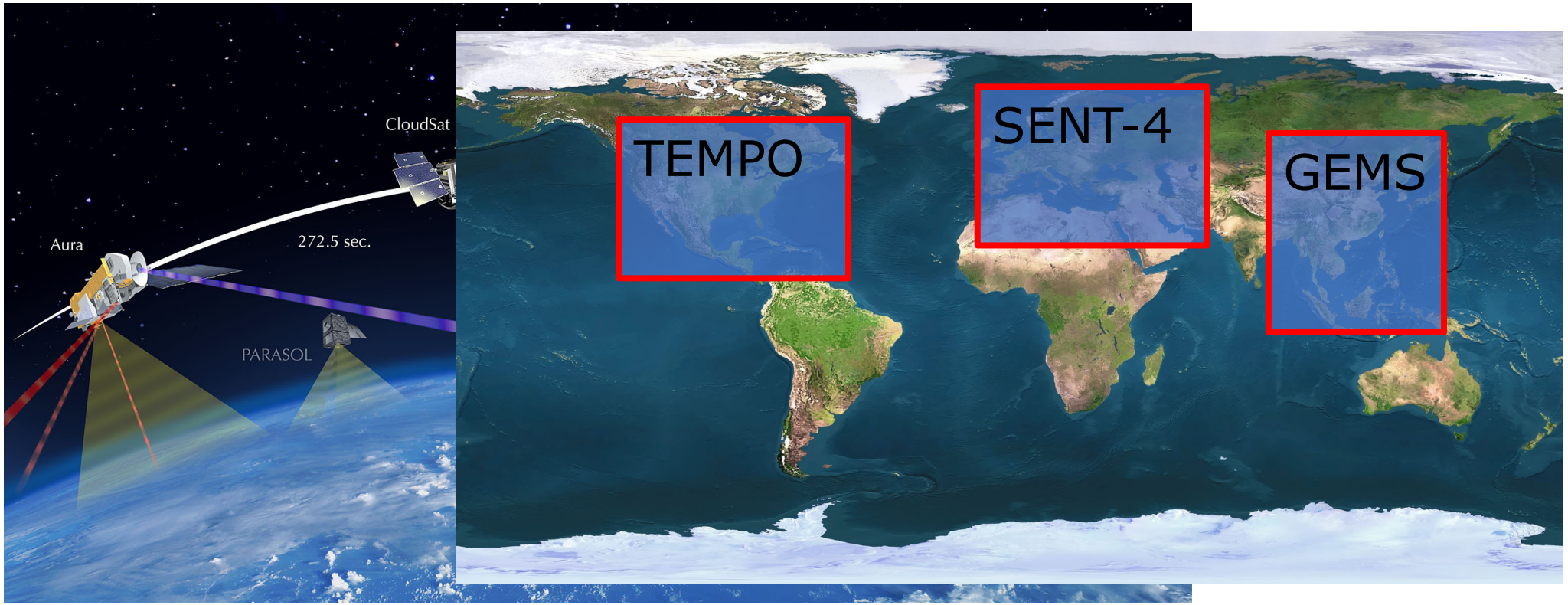




Elguindi et al, 2020

Data assimilation and inverse modeling

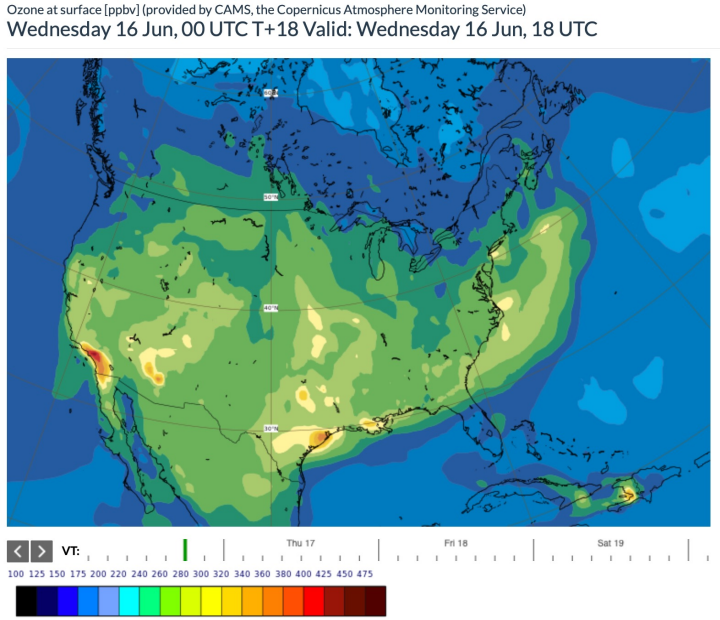
We have large networks of observations of atmospheric **composition** (aerosols, short and long-lived gases) from recent and future satellites



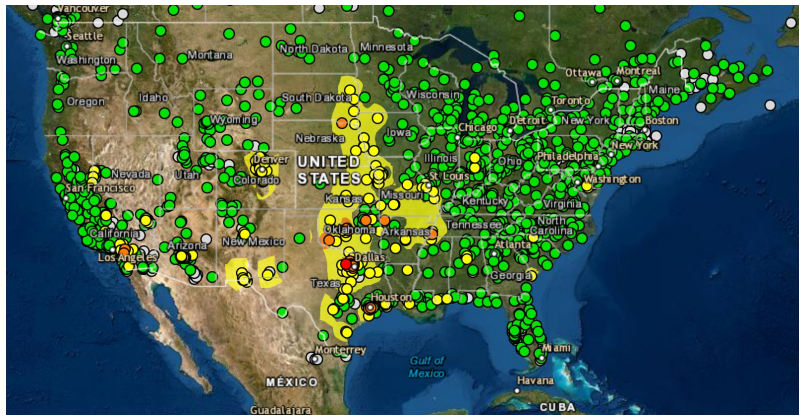
What do composition measurements tell us about the emissions?
How does the state of a system contain information about the boundary conditions?

Operational chemical data assimilation and AQ forecasting

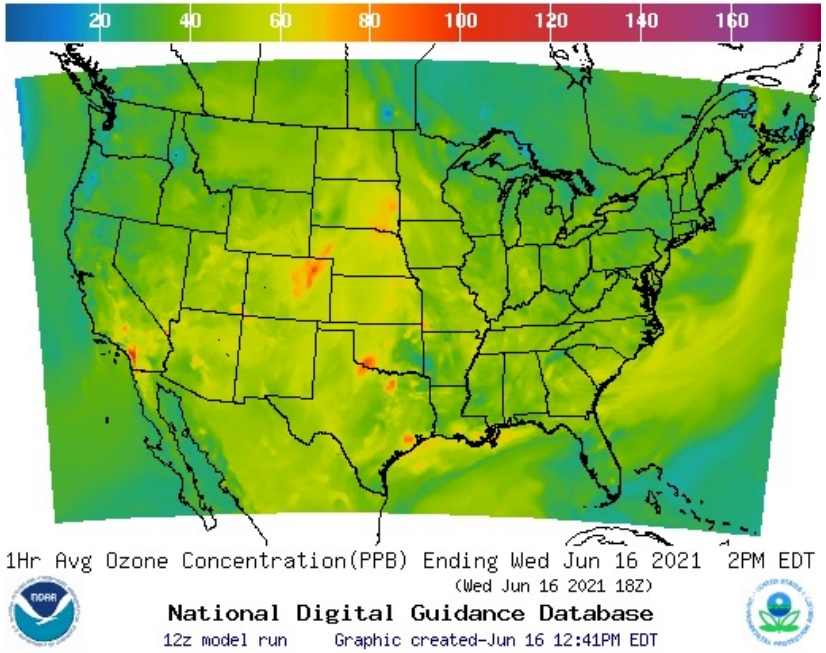
Europe: ECMWF



AirNow 3 pm EDT



US: NAQFC

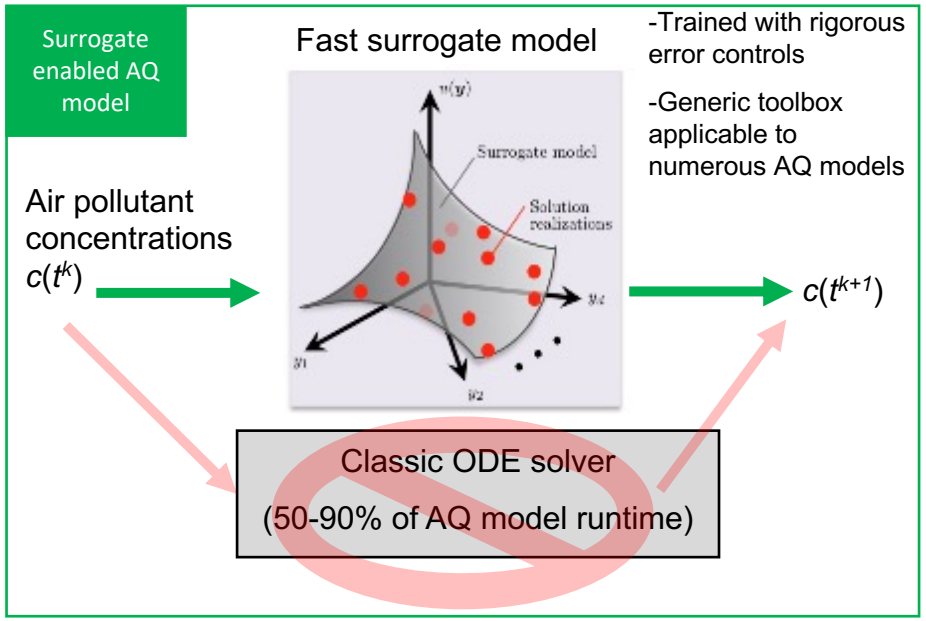


- routine assimilation of trace-gas satellite data
- use overly simplified representation of chemistry

- assimilation for fires, smoke,
- no routine assimilation of trace-gas satellite data

Computational cost prevents system for assimilating full suite of data using high-fidelity AQ model

Background: computational challenges for AQ models



The ODE solver can be replaced with a more efficient surrogate model to reduce computational cost

Previous machine-learning work in atmospheric models:

- Ozone ensemble forecast with machine learning algorithms [Mallet et al. (2009)]
- Using machine learning to build temperature-based ozone parameterizations for climate sensitivity simulations [Nowack et al. (2018)]
- Orders-of-magnitude speedup in atmospheric chemistry modeling through neural network-based emulation [Kelp et al. (2018)]
- Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem … [Keller & Evans (2019)]
- Improving the prediction of an atmospheric chemistry transport model using gradient boosted regression trees [Ivatt & Evans (2019)]

Objectives

Objectives and **Information Technology** (bold) :

- Develop, test, and deliver a **surrogate model for chemistry in GEOS-Chem**
- Generalize **surrogate model generation** procedure within a **software toolbox**
- Demonstrate benefits of **surrogate-based AQ modeling framework for chemical data assimilation** of geostationary observations of atmospheric composition

Science goals:

- Develop new techniques for surrogate modeling of high-dimensional, non-linear, large-scale dynamical chemical systems
- Improve O_3 forecasting through assimilating NO_2 observations from geostationary remote sensing measurements.

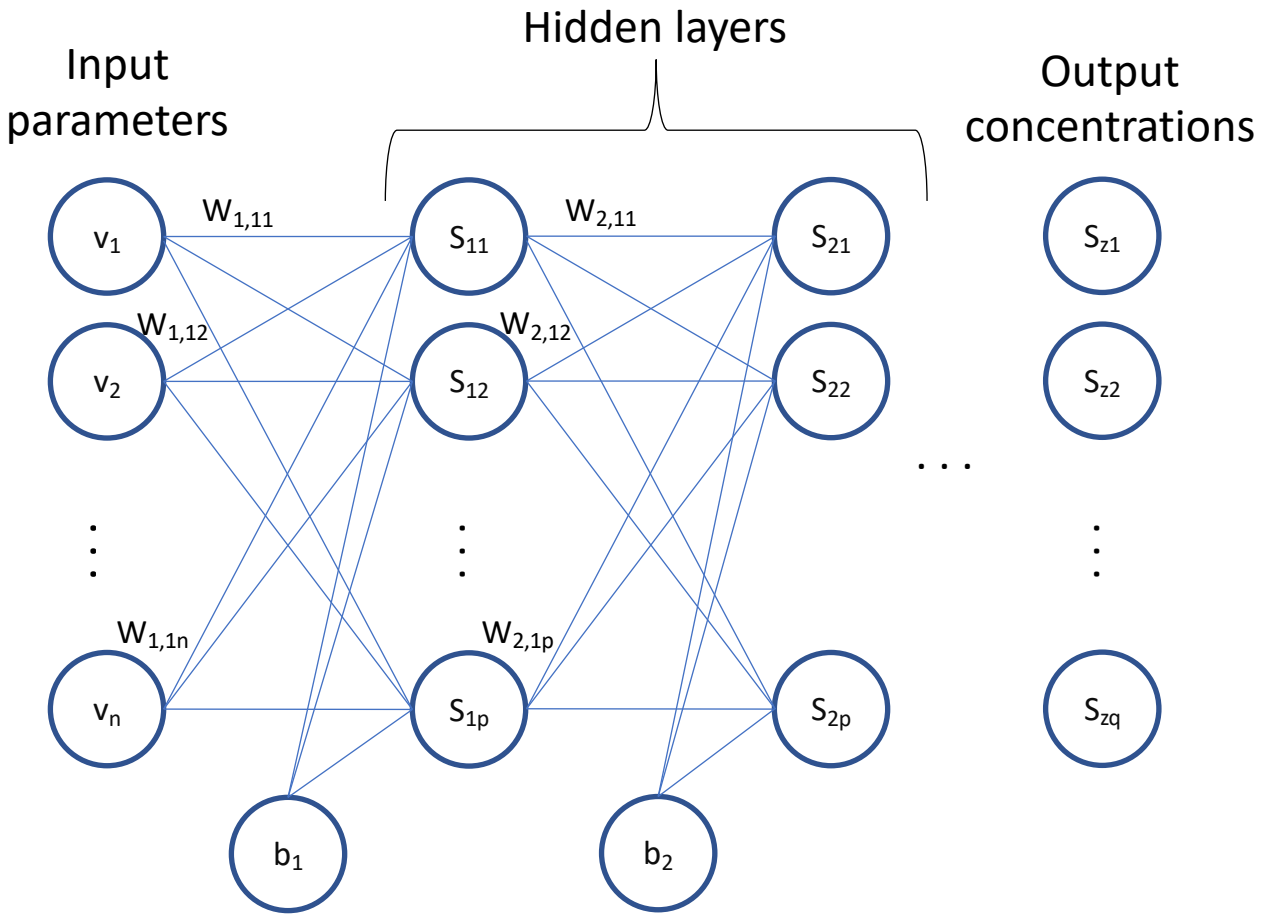
Partners:

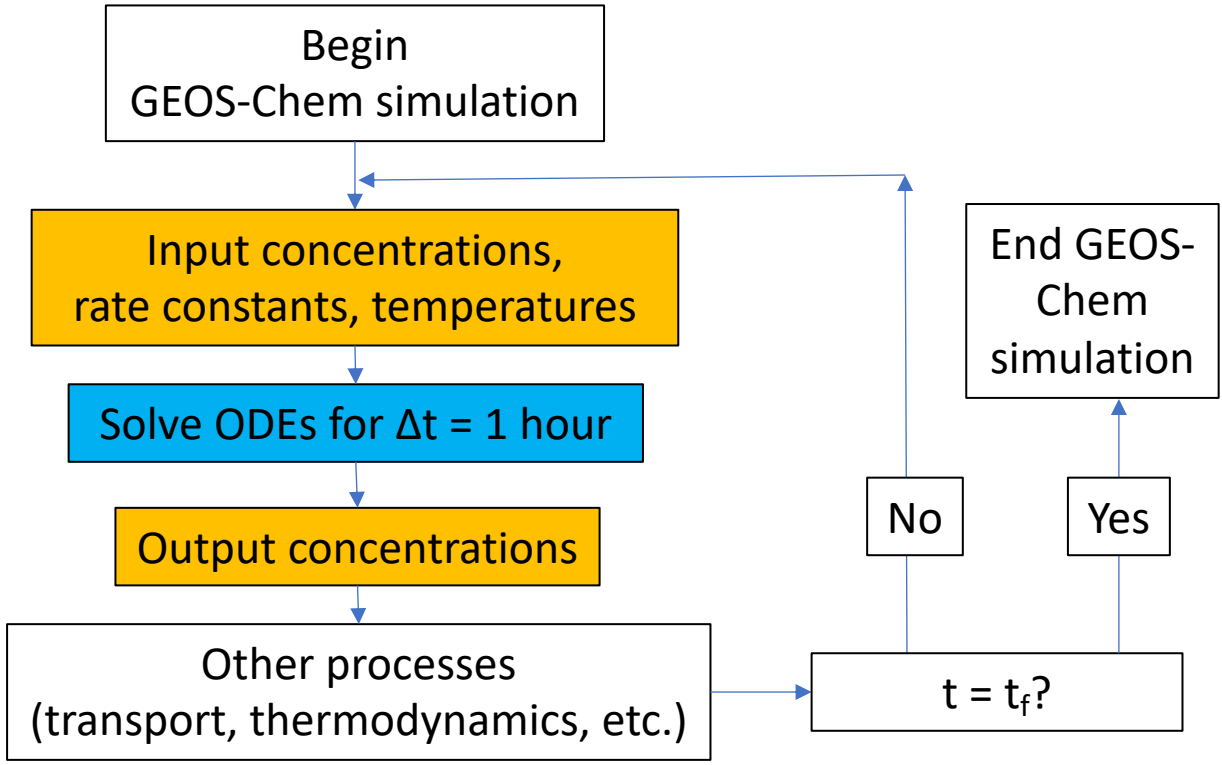
- Nicolas Bousserez, ECMWF

Funding:

- NASA AIST AIST-18-0072

Overview of deep neural network & training data generation





Sampling for training data to generate the surrogate model was taken randomly from hourly global GEOS-Chem 2x2.5 resolution simulations.

Parameters for training the neural network

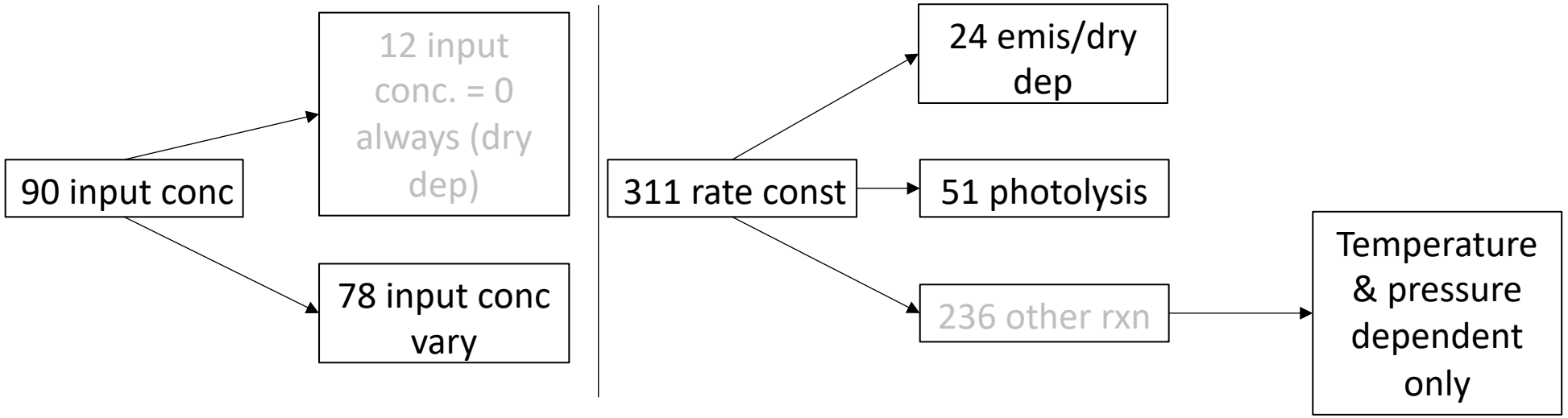
Currently, the ODE solver (for GEOS-Chem full chemistry) takes as inputs:

- 90 chemical species with variable concentrations
- 311 rate constants, each representing one reaction or other process

And outputs:

- 90 chemical species with variable concentrations

Of the 401 inputs to the ODE solver,

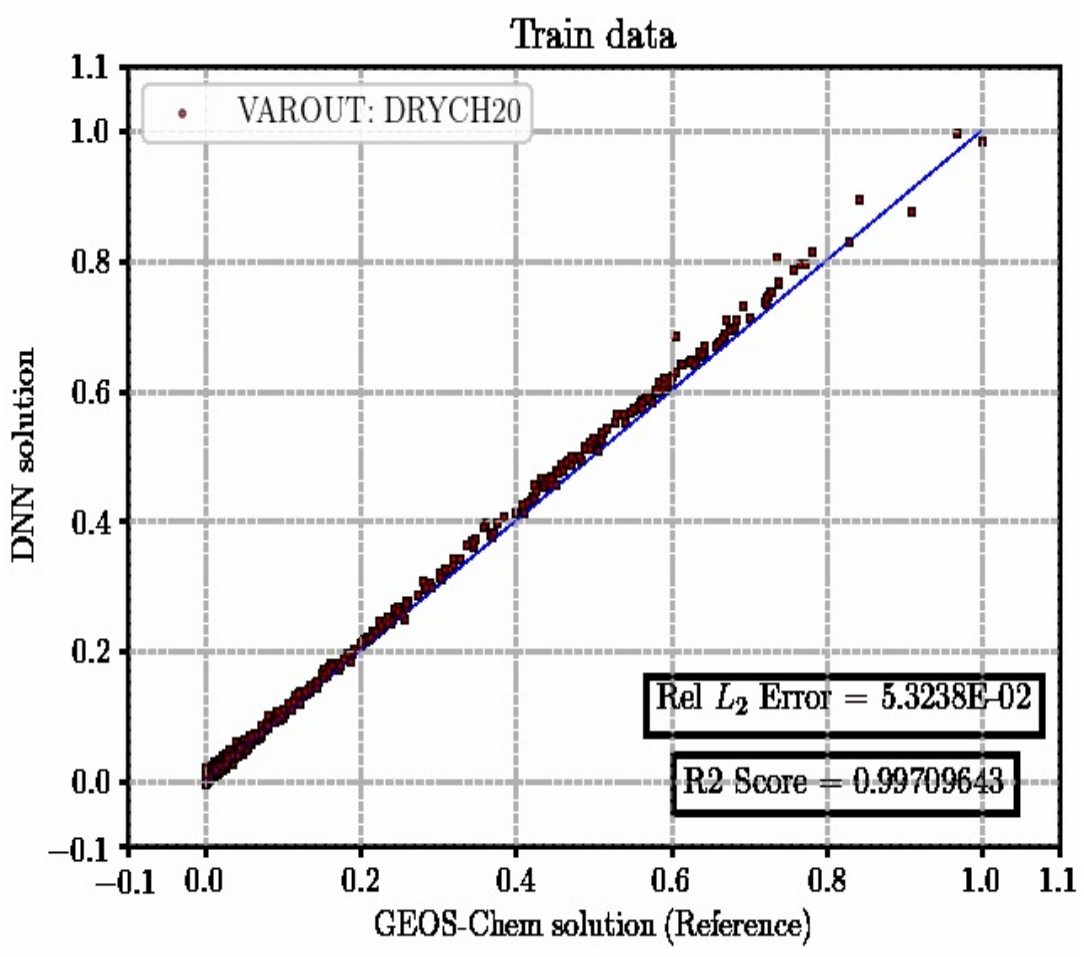


For the surrogate model training, we have reduced the number of variables for each grid cell from 491 to 245:

- 78 input concentrations, 90 output concentrations, 75 rate constants, 1 temperature, 1 pressure

Satisfying R2 measures of the 90 surrogate models

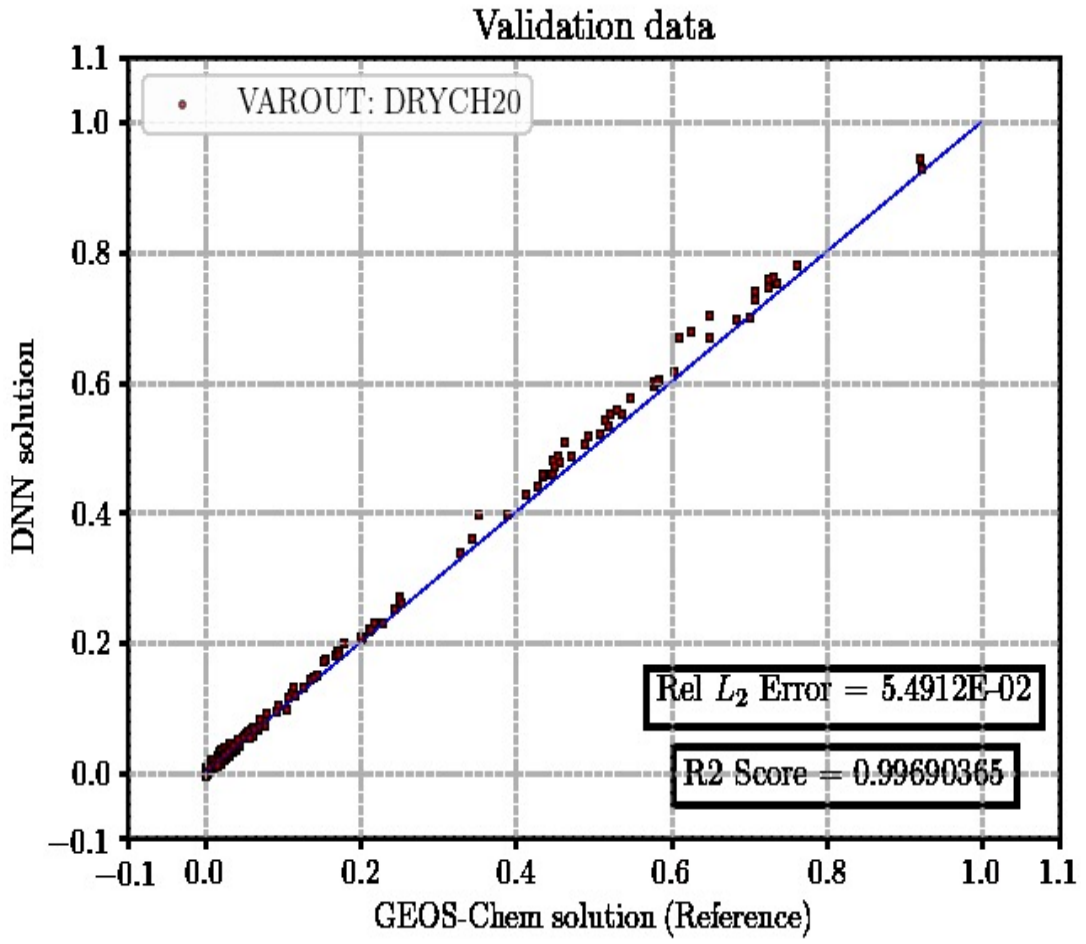
- **DNN model (Ver 0)**
  (1) Independent surrogate models for respective chemical species
  (2) R_2 values > 0.98 for all surrogate models



Samples from GEOS-Chem global 2x2.5

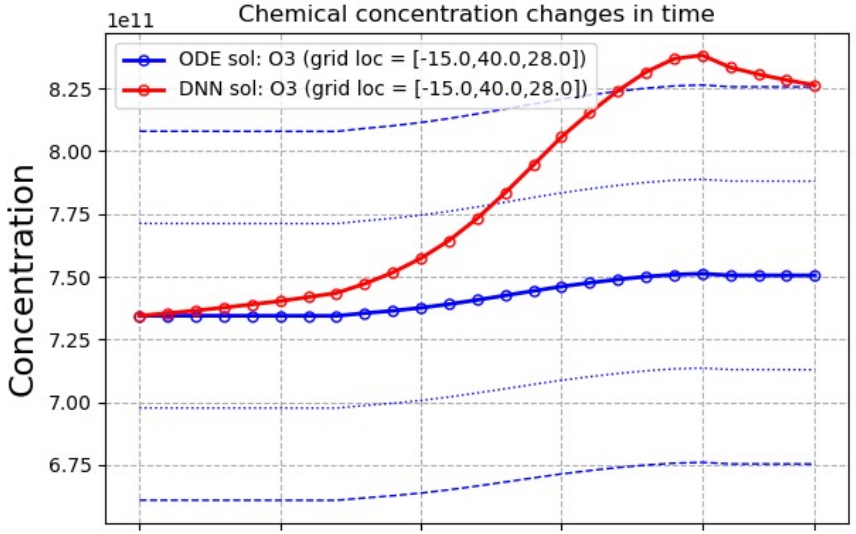
Satisfying R2 measures of the 90 surrogate models

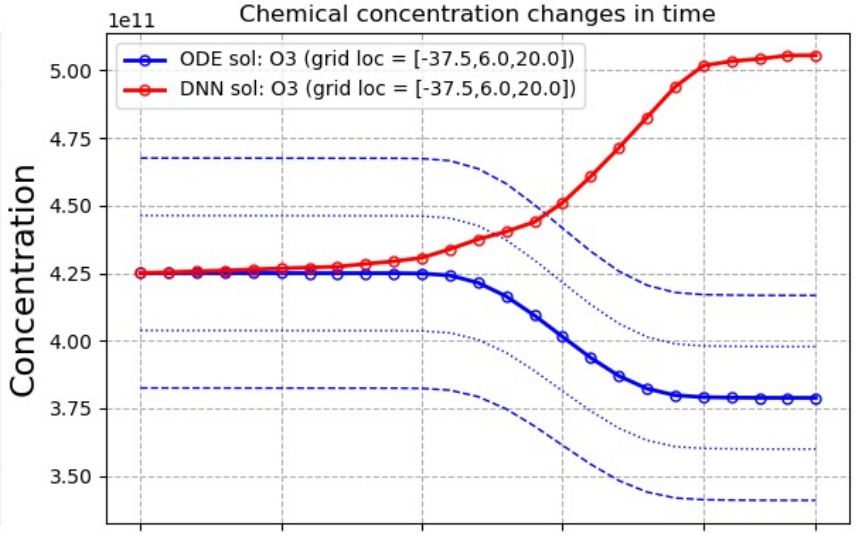
- **DNN model (Ver 0)**
  (1) Independent surrogate models for respective chemical species
  (2) R_2 values > 0.98 for all surrogate models

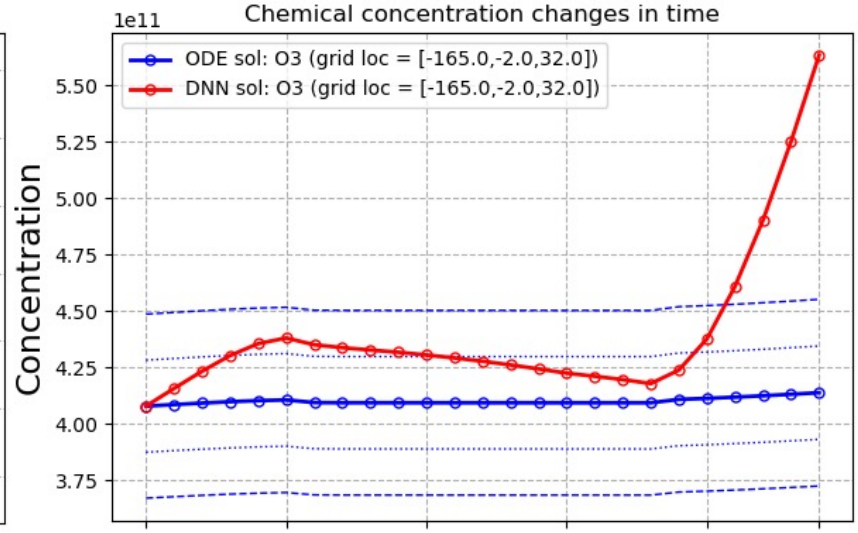


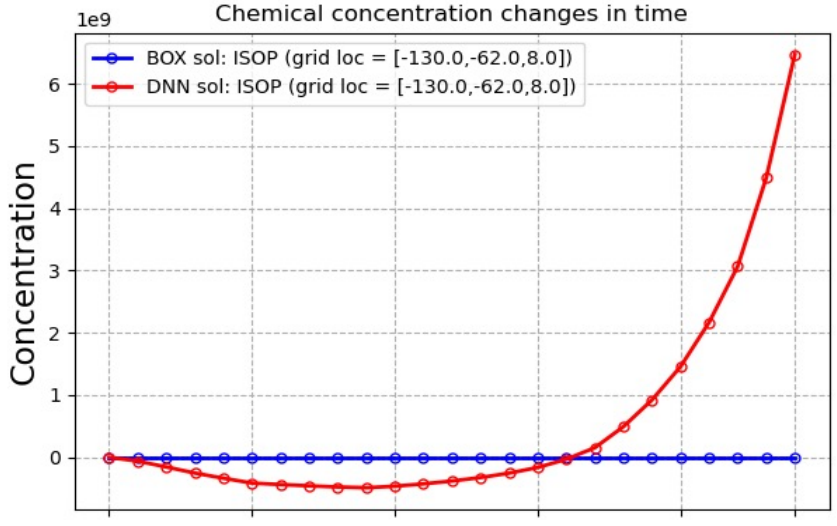
Samples from GEOS-Chem global 2x2.5

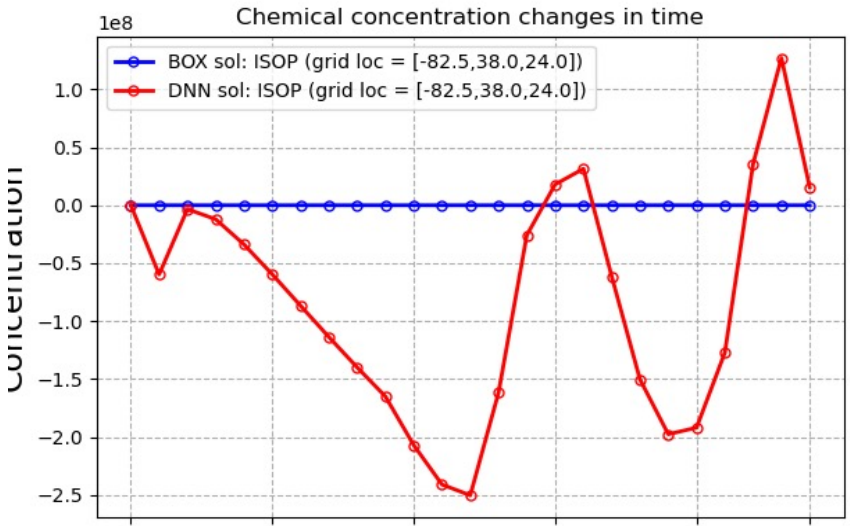
Time-transient solutions (top: O3, bottom: ISOP)

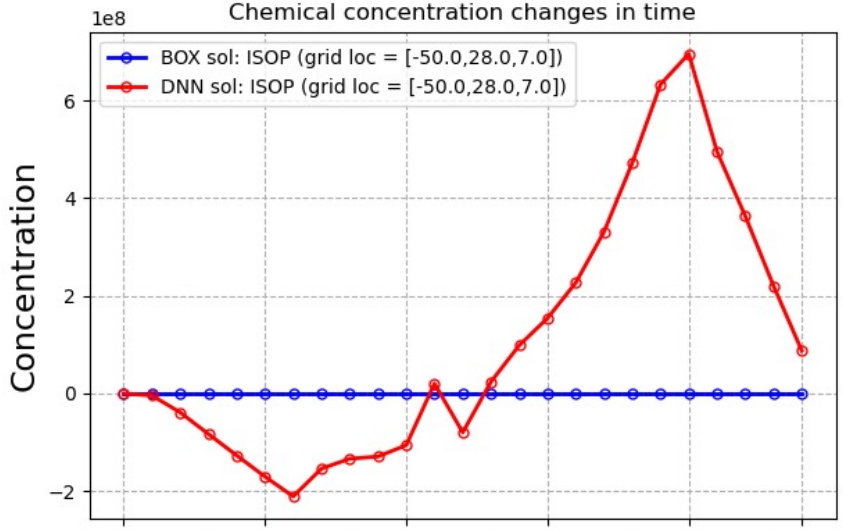






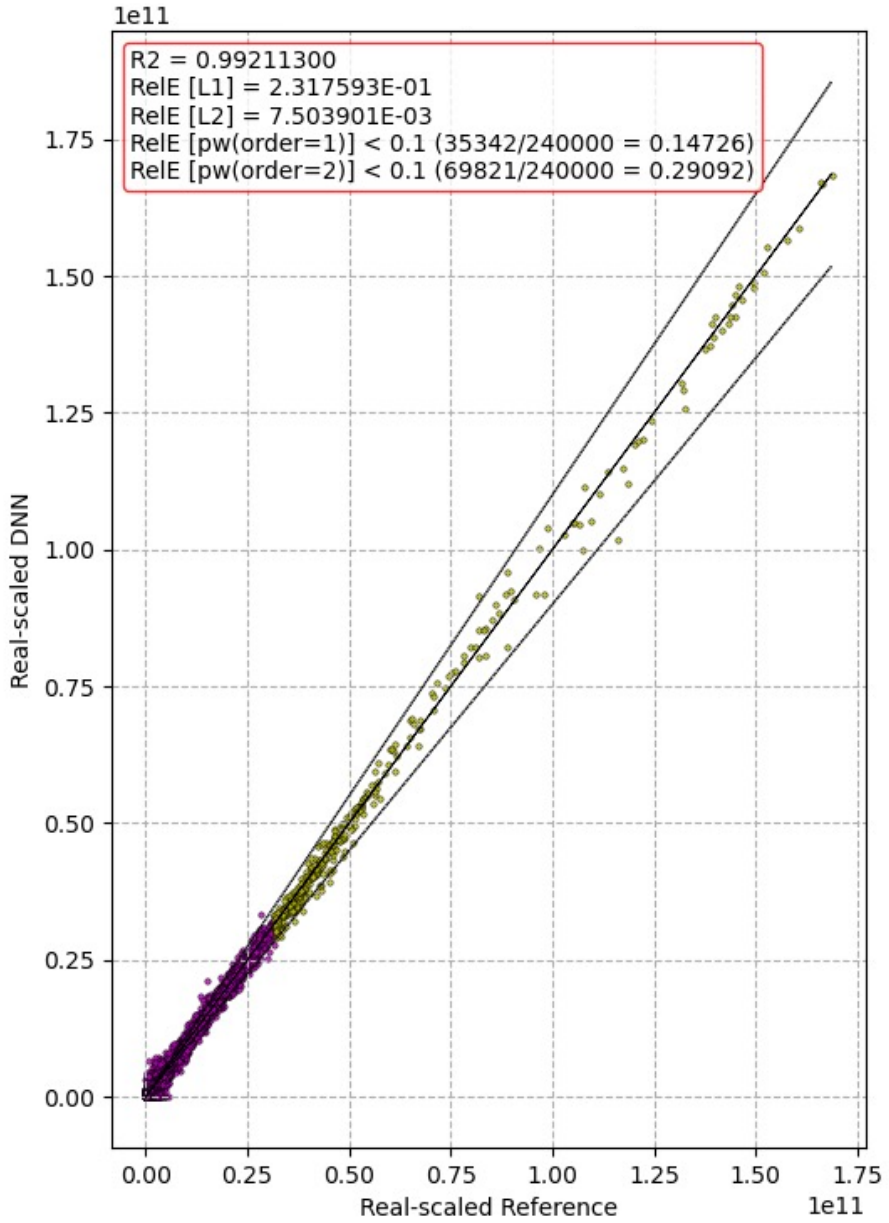


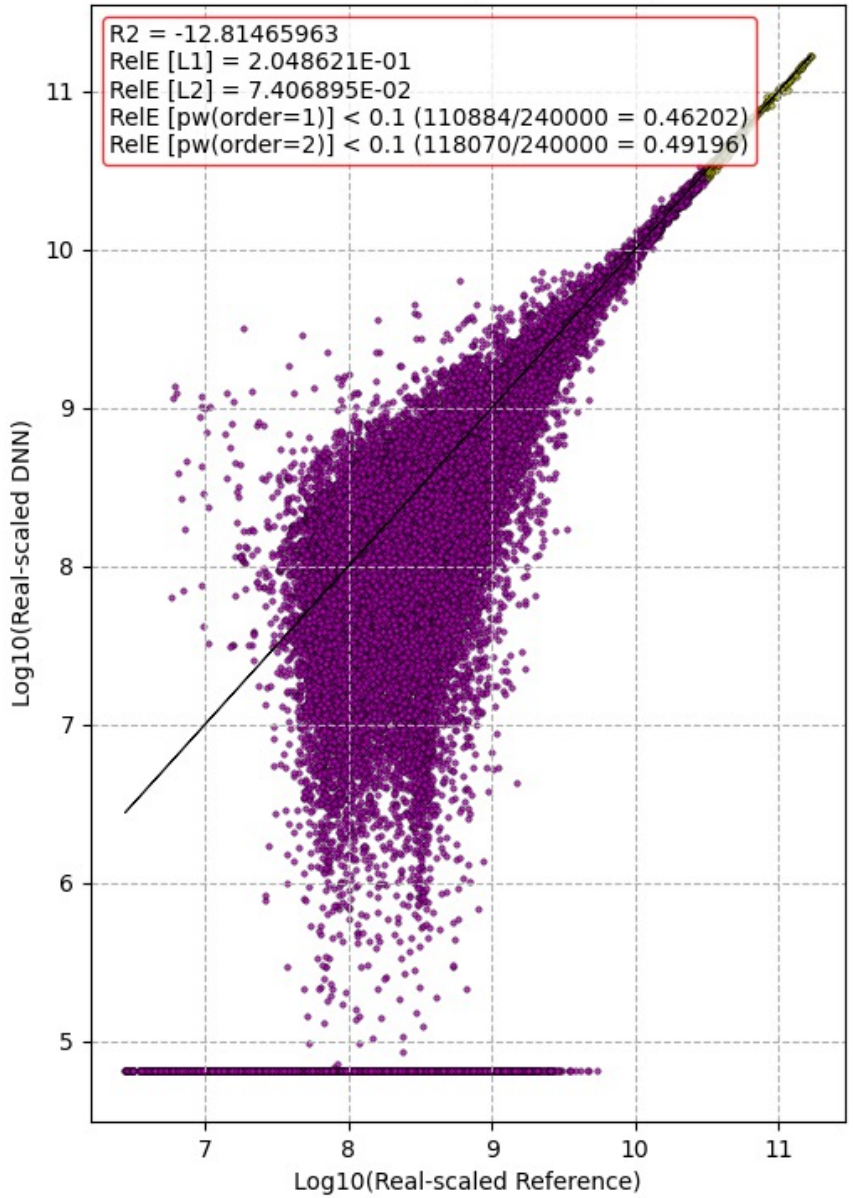




Difference in real-scaled and log-scaled comparisons

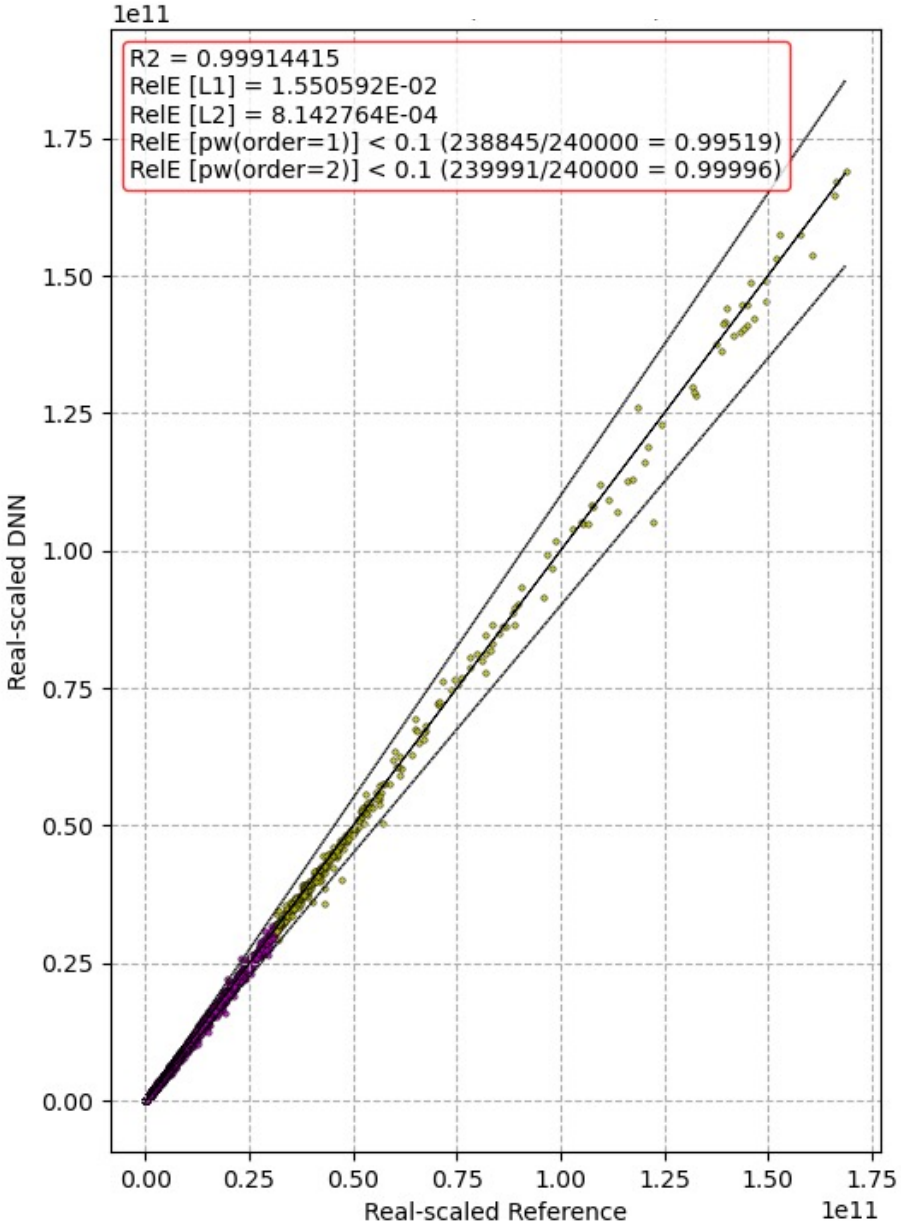
- NO2

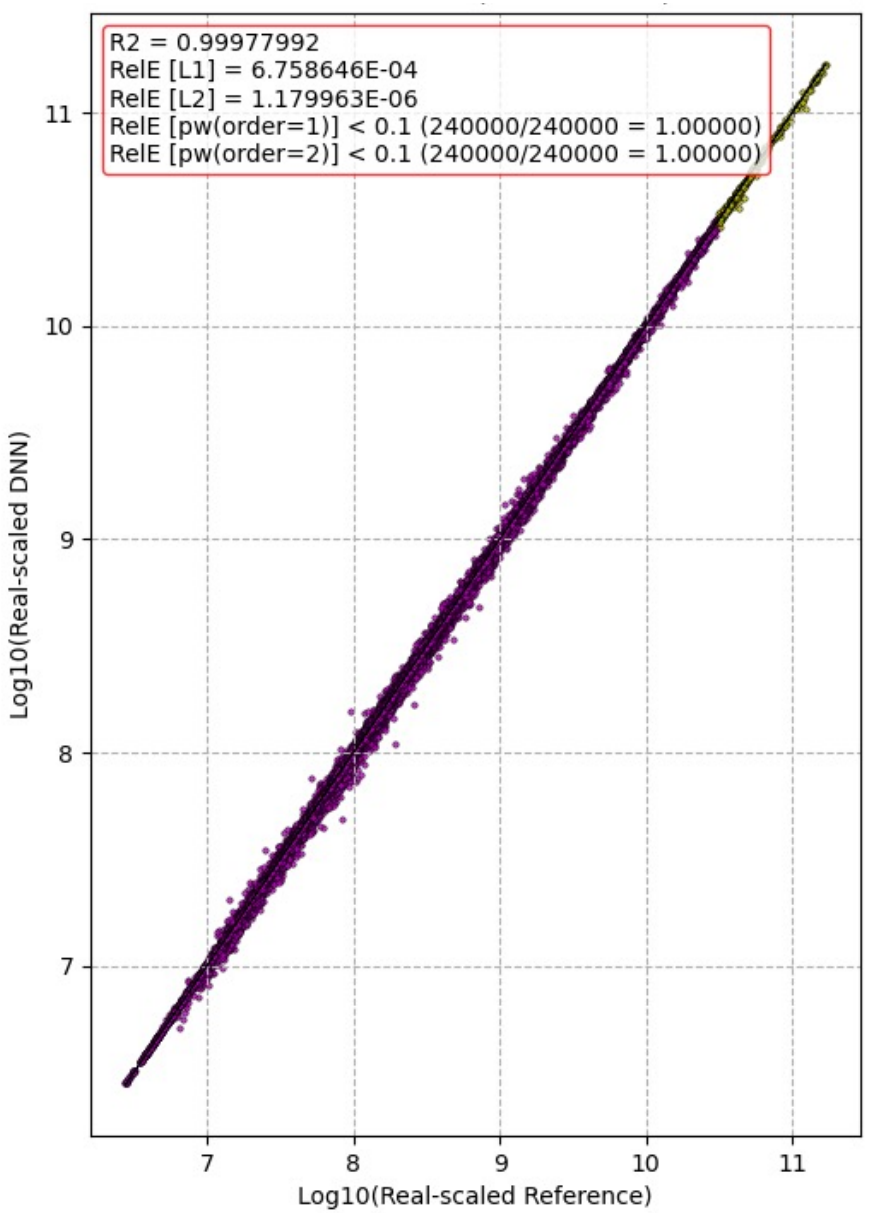




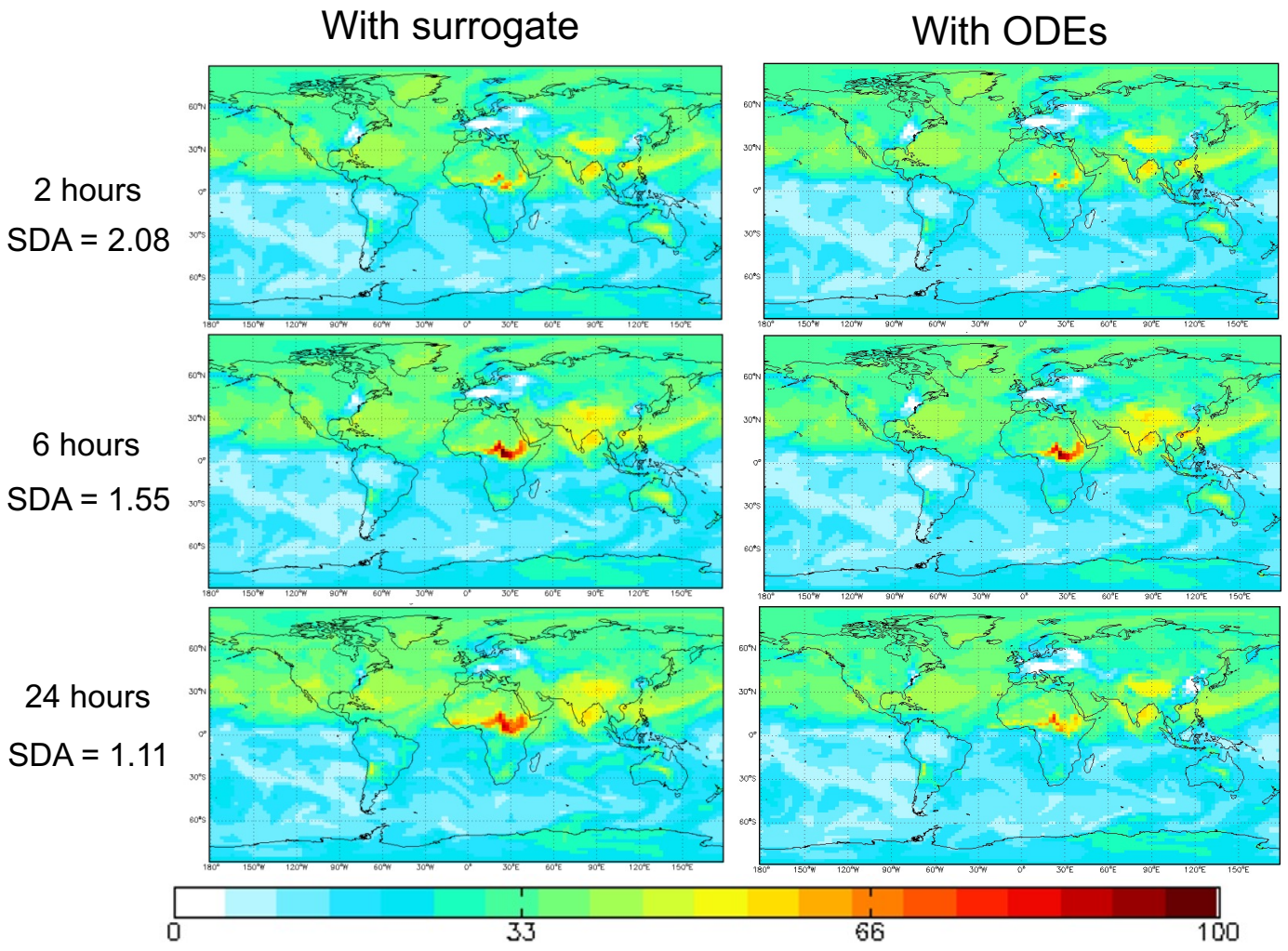
Improvement from log-scaled training

- NO2





GEOS-Chem simulations of O_3 over 24 hours



90 separate surrogates are used here, as previously described (one per chemical species)

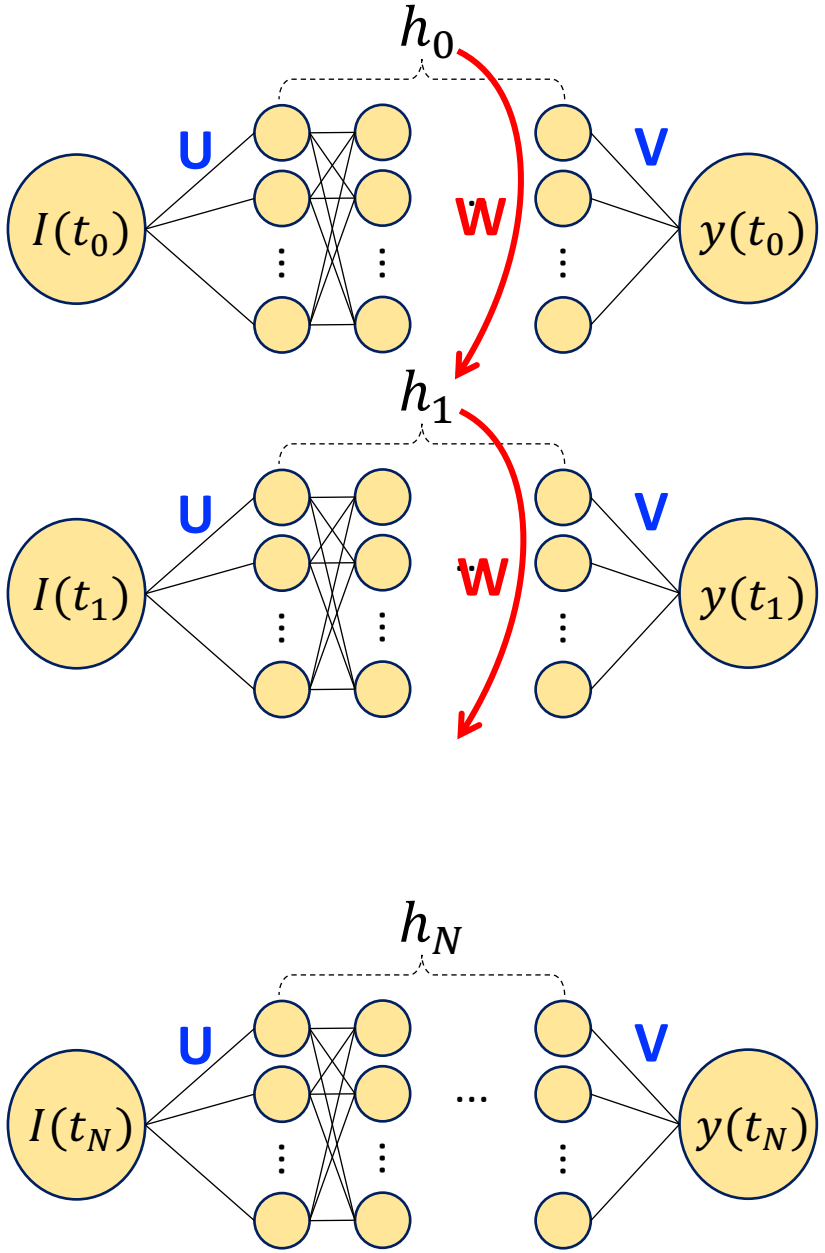
Significant digits of accuracy (SDA) - grid cell modified root mean square norm [Sandu et al. (1997), Henze et al. (2007)]

$$SDA_{O3} = -log_{10}\left(\sqrt{\frac{1}{|\theta_{O3}|}\sum_{j\in\theta_{O3}}\left|\frac{c_{O3,ODE,j}-c_{O3,surr,j}}{c_{O3,ODE,j}}\right|^2}\right)$$

θ_{O3} = # of grid cells with c_{O3} > 1e11 molec/mL

Recurrent Neural Network (RNN)

University of Colorado Boulder



- **the *j*-th data sample**

$$I(t_0)_j = \{\tau(t_0, x_j), p(t_0, x_j), \boldsymbol{C}(t_0, x_j), \boldsymbol{k}(t_0, x_j)\} \qquad \boldsymbol{C}(t_1, x_j) = y(t_0)_j$$

$$I(t_1)_j = \{\tau(t_1, x_j), p(t_1, x_j), \boldsymbol{C}(t_1, x_j), \boldsymbol{k}(t_1, x_j)\} \qquad \boldsymbol{C}(t_2, x_j) = y(t_1)_j$$

$$I(t_N)_j = \{\tau(t_N, x_j), p(t_N, x_j), \boldsymbol{C}(\boldsymbol{t_N}, x_j), \boldsymbol{k}(t_N, x_j)\} \qquad \boldsymbol{C}(\boldsymbol{t_{N+1}}, x_j) = y(t_N)_j$$

- **Input (I) and Output (y) relationships**

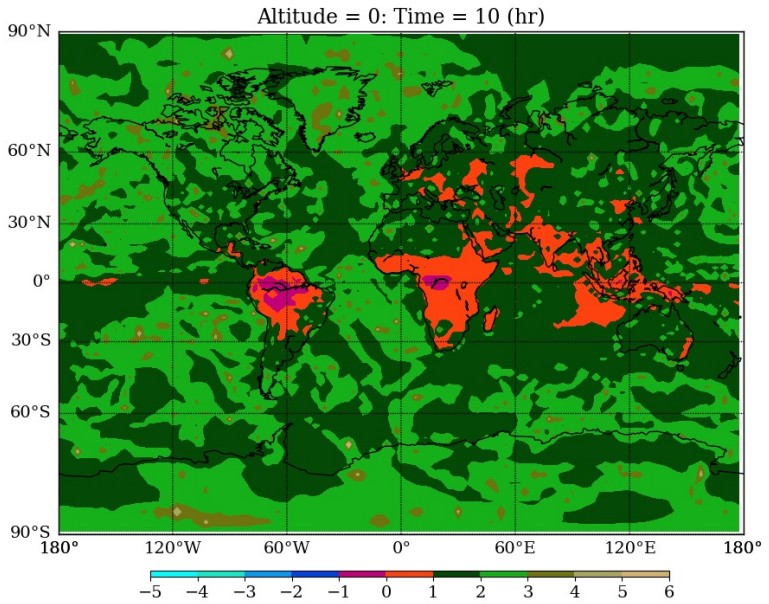
$$y(t_0) = \varphi(\boldsymbol{V}\, h_0), \qquad h_0 = \sigma(\mathbf{U}\, I(t_0))$$
$$y(t_1) = \varphi(\boldsymbol{V}\, h_1), \qquad h_1 = \sigma(\mathbf{U}\, I(t_1) + \boldsymbol{W}\, h_0)$$
$$\vdots \qquad\qquad \vdots$$
$$y(t_N) = \varphi(\boldsymbol{V}\, h_N), \qquad h_N = \sigma(\mathbf{U}\, I(t_N) + \boldsymbol{W}\, h_{N-1})$$

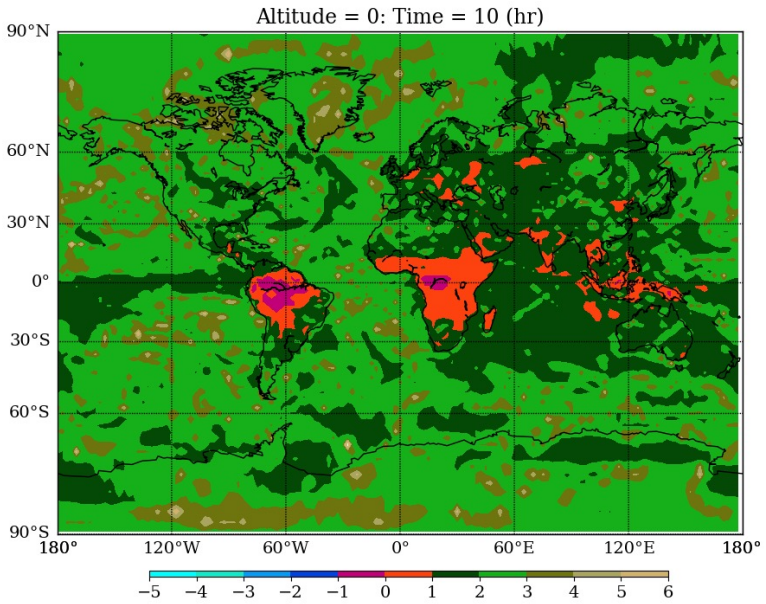
* **U, V, W**: independent of t

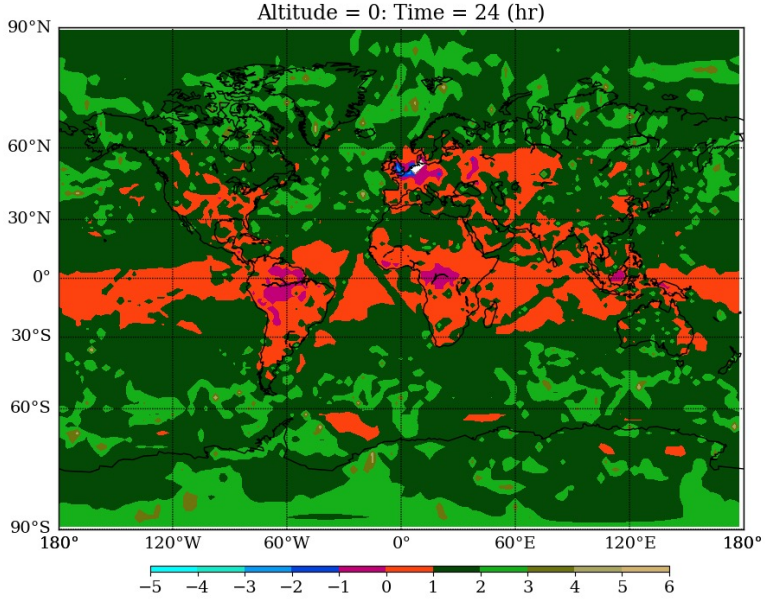
Improvement of O3 prediction from RNN network

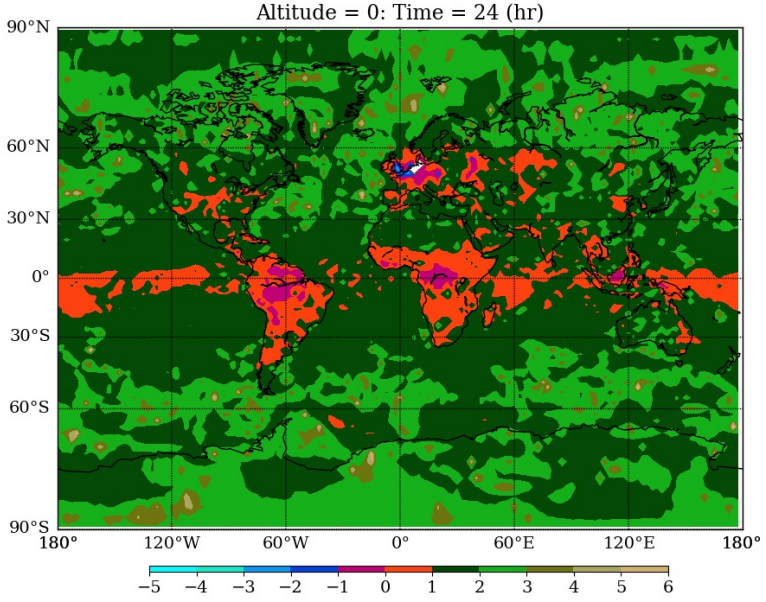
University of Colorado Boulder

- Previous model (left)
  vs.
  RNN model (right)









Summary

- Surrogate model development
  - Ensemble of DDN models (one for each individual concentration)
  - Tested / developed in box-model R&D using samples from 3D model

- 3D implementation:
  - Initial implementation (Ver 0) reasonable for a few hrs followed by error growth
  - Now adding flexibility to accommodate multiple surrogate model versions

- Surrogate model updates for improved accuracy:
  - Non-negative constraints
  - Chemical and physical regimes
  - Log-scaling
  - RNN

- Next steps:
  - Explore additional ideas for enhanced stability / accuracy
  - Apply to chemical data assimilation with GEOS-Chem
  - Apply to other models: collect samples from CAMS model (ECMWF collaborator)
  - Towards implementation of surrogate model generation within an Air Quality Analytic Center (AIST supplement, PI Thomas Huang, JPL)